Systematic Cardiovascular Disorder Identification Using Machine Learning Algorithms

***S Subbulakshmi [1], K V Adarsh[2]**

[1]Department of Computer Science and Applications

[2]Department of Computer Science and Engineering

Amrita Vishwa Vidyapeetham,Amritapuri,India

[2]kvenkataadarsh@am.students.amrita.edu

*Corresponding email: subbulakshmis@am.amrita.edu

Abstract: One of the major causes of world's high death rate is heart disease. Massive volumes of data related to clinical trials and analysis are stored on biomedical equipment and other systems in the hospital. As a result, understanding the data associated with heart disease is crucial for enhancing prediction accuracy. In this study, the performance of models created with machine learning classification algorithms and standardized characteristics derived with various feature selection approaches was tested experimentally. This study investigated the possibilities of classification approaches, notably decision trees, K-Nearest Neighbour (KNN), Support Vector Machine (SVM) and Random Forest (RF), for the prediction of heart disease. Medical features of individualslike age, gender, blood pressure, fasting blood sugar, and the type of chest discomfort can be used to predict an individual's risk of getting heart disease. Consequently, medical community is flourishing. when compared to previous methods, classification-based approaches have shown to be highly effective and accurate. This research compares many ways for predicting heart issues. The results of this study will help researchers better grasp the present approaches for developing heart disease prediction models. This research presents the findings of a study of key machine learning algorithms that could be utilized to develop a highly accurate and efficient prediction model to help physicians reduce the number of heart disease-related deaths.

***Keywords**:*Heart disease, Machine Learning, Prediction, Classification techniques, SVM, KNN, Random Forest, High accuracy.

1. Introduction

Cardiac auscultation is a widely non-invasive and cost-effective method for early detection of congestive heart, valvular heart problems, failure, and basic heart abnormalities [1]. Successful cardiac auscultation, on the other hand, necessitates the hiring of trained doctors, which are in short supply in

rural areas and low-income nations worldwide. In addition, being a physician is a difficult, time-consuming, and subjective job. As a result, machine learning-based automated heart sound classification systems may have a substantial influence on the early diagnosis of cardiac illnesses [2]. The heart is a vital core-muscle organ in every living human, pumping blood to other organs of the body via blood arteries of circulatory system. The proper functioning of heart is most critical for the survival of all individuals.

Coronary diseases, cardiomyopathies, angina pectoris, congenital, and other heart and artery problems are all instances of heart problems. A growing proportion of young people are being diagnosed with cardiac disorders. Aside from the components, it is considered that common risk factors are caused by a range of lifestyle choices such as physical dormancy, unhealthy eating habits, and obesity [3,4].

It is difficult to diagnose heart illness manually based on risk factors. Conversely, machine learning categorization algorithms based on existing data can anticipate the illness [5]. This study intends to emphasise on the significance of researchers' work, along with the dataset used to predict heart disease from risk indicators, and the machine learning classification methods which used the selected dataset [6].

This article is organised as chapter 2 Related Works which elaborates the existing research work done in this domain, chapter3 Proposed Methodology describes the implementation aspects of the above system which includes pre-processing, essential feature selection and machine learning techniques used for prediction. Chapter 4 Results and Discussion explains the accuracy calculation with is results used analysing the most promising model and Chapter 5 Conclusion summarises the work done in this heart disease identification system with possible scope for future enhancements.

2. Related Works

Occurrence of Heart Disease Prediction and Analysis proposed byChalaBeyene et al. [7] endorsed the use of various data mining methods. Its primary objective is to forecast the onset of cardiac sickness so that an early automated diagnosis and prognosis may be delivered in a timely manner. The proposed technique is equally important in a healthcare organization with staff that lack extra skills and abilities. It analyses a range of medical parameters which includes blood sugar, age, heart rateand gender, to determine whether a person has cardiac illness. WEKA software with rich set of application program interface (AIP)is used to analyse the dataset. Latha et al. in their research work [8] uses ensemble classification with feature selection methods to generate a model that foresees the risk of heart disease. According to the findings, ensemble techniques like bagging and boosting plays a significant role in enhancing the prediction performance of bad classifiers and performs well in detecting heart disease

risk. Moreover, feature selection approach was used to boost the performance even more, resulting in a considerable improvement in prediction accuracy.

Jagdeep Singh et al. [9] have used cardiac datasets to predict heart disease using a variety of correlations and classification techniques. They built a decision system for predicting cardiac disease using classification associative rules (CARs) in the WEKA framework, employing a hybrid technique that included classification and association.

Mathan et al. [10] used innovative information quality to examine augury frameworks for cardiac disease. In this paper, they introduced a novel technique of categorising based on decision trees, which yields accurate findings that differ significantly from earlier calculations. They used the prior calculations' voting mechanism. Voting mechanism of discretization methodologies are used to create more precise decision trees with highprediction accuracy and sensitivity. During execution, research work analysed the outcomes by applying variouscombinations of instructions to several types of decision trees and executing the appropriate decision tree techniques to attain high accuracy and sensitivity.

In this research work we used dataset from UCI repository for prediction of heart disease. The initial step of implementation includescleansing of dataset to enable accurate prediction results using different pre-processing techniques which is followed by principal feature selection process to selecting most promising features. Then different machine learning algorithms [11] are used for creating models to identifycardiovascular disorders based on the parameters in the dataset. Accuracy of the models is assessed to conclude with the most promising model. Comparative study of prediction results is done based on different factors like specificity, sensitivity, and accuracy. Finally, model with highest accuracy is finalised and it used for future predictions when a new patient data is feed into the

3. Proposed Methodology

It is highly alarming that even young people who seems to be so healthy are affected by sudden or silent heart attacks. Only old people go for periodical check-up others are busy with their day-to-day cores, and they fail to take care of their health. The proposed work thrives to collect some basic features from different individuals pertaining to heart disease and uses the same to predict whether the person have the chance of getting heart disease or attack in near future. Thus, cardiovascular disorder identification system would enable users keep track of theirhealth in the perspective of heart related issues.Flow of the proposed system with all its major components is shown in the following figure 1.

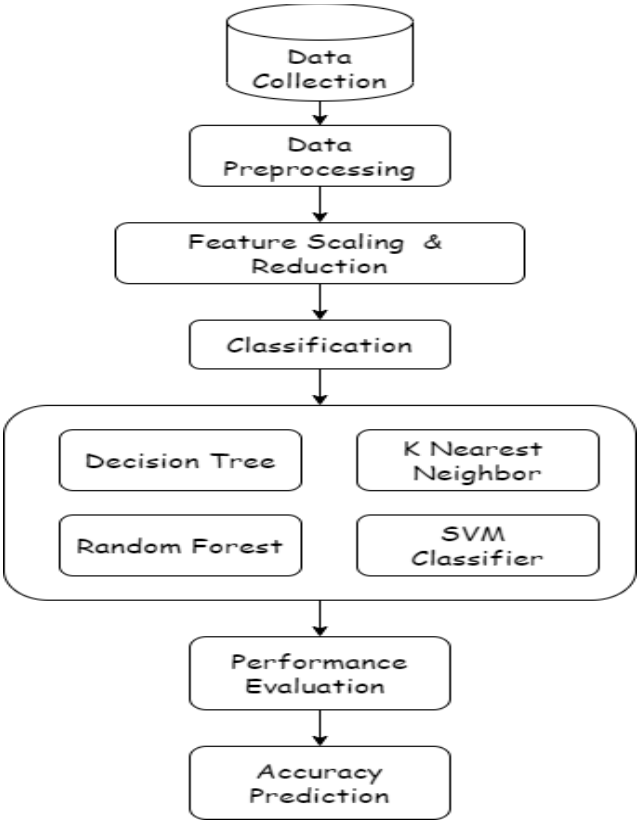

Figure 1: Proposed Methodology.

A) Input Dataset taken from UCI Repository

The open-source dataset registry of the UCI dataset is used in the analysis. It contains a variety of disease-related databases. This model makes use of the UCI dataset on heart disease. The Cleveland Heart Disease dataset from the UCI repository is used in our system. The UCI Dataset for heart illness study is used to create a model that successfully predicted heart illness by combining different machine learning classifier algorithms. The data collection contains both related and irrelevantinformation. Relevant information has direct consequences on the output field and irrelevant data does not much impact on the prediction results. Obviously, useful data is chosen for identifying promising features with appropriate methods. Data pre-processing [12] isperformed on the data set for null valueremoval, normalizing the data values etc. It is very clear from the previous studies that quality of the dataset plays vital role in the prediction results of all machine learning algorithms.

B) Data Pre-Processing

The dataset on heart disease is pre-processed once a large number of records have been gathered. The dataset contains 3030 patient records, with missing data points. After identifying those missing entries from the dataset, suitable approach for handling missing values is determined. Missing value

management is a data preparation approach for constructing a smooth dataset. Consequently, it began by searching the dataset for missing values. missing values can be handled in a variety of ways, including outright rejecting them, replacing them with any numeric value, swapping them with the most often occurring value for that property, and so on. They should be replaced with the mean value of the characteristic.Around 0.3 % variables are missing from the heart disease dataset which used in this study. The attribute's mean or median values are used to represent null values.

C) Feature Selection and Reduction

Using two factors relating to age and sex, the patient's personal information is recognized from among the 13 items in the data set. The following 11 characteristics are highly graded because they include critical treatment information. Clinical data is vital for diagnosing and identifying the severity of heart illness.

Advantages

i) Improved sensitivity in detecting heart disease

ii) Uses a random forest technique and feature selection to deal with the most difficult (massive) amount of data.

iii) Reduce the amount of time it takes for doctors to complete their tasks.

iv) Patients can afford it.

Implementationis tested with and without feature extraction toassess its overall impact in the prediction accuracy of the system. The main goal of this process [13] is to identify the most important components of cardiac issues. Furthermore, by deleting dataset characteristics, feature extraction facilitatesthe construction of accurate model by rejecting under-representing less relevant characteristics, hence lowering time spending in training the dataset and for improving the learning process.

D) Machine Learning Algorithms

It is an emerging subfield of artificial intelligence, with the primary objective to develop systems that can learn, and estimate based on historic data. It finalises a model by implementing machine learning algorithms [14] on a training dataset. Model for this system anticipates heart disease using the sufficient input data. Four machine algorithms are described in this paper for the prediction or classification of heart disease based on medical data characteristics: SVM, KNN, DT, and RF. The following is a technique for anticipating heart disease:

i) The heart infection data package is retrieved from the UCI ML repository.

ii) In the detailed packages, numerical values must be substituted for various NaN values. This method is used during the pre-processing step.

iii) Our data package is divided into training and testing data for validation.

iv) Finally, several techniques are employed to train training data, and the learned model is used to categorise testing data.

Decision Tree (DT)

It builds models by using machine learning to discover hidden patterns in the input dataset. It provides reasonable forecasts [15] for fresh datasets. The dataset has been sanitised and any missing values have been filled in. After anticipating heart disease using the new input data, the model is evaluated for accuracy. the optimal split for each node This approach is used to determine the information gain, index, and dividend growth rate.

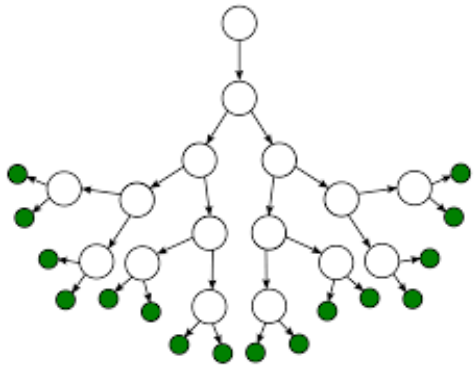

Figure 2: Decision Tree Structure

Following the determination of the entropy of each feature, the dataset is partitioned using parameters with the highest data gain or the lowest entropy. The remaining components are recursively employed to carry out the next two steps.

$$Entropy(E) = \sum -q_k \log_2 q_k$$

A decision tree-based classifier with and without voting techniques of a discretization approach has 89.9 percent sensitivity and 86.5 percent specificity, respectively. The decision tree had the lowest accuracy of % in, but the authors increased it to 93.6 % by combining it with a boosting method.

Support Vector Machine (SVM)

It is a supervised machine learning approach for categorising data and spreading it throughout a subspace. As shown in Figure 3, a hyperplane is a line which is used to divide a plane into two halves

in two-dimensional space. An SVM technique [16] is used to upsurge the margin, which indicates the distance between hyperplane plane and the two nearest data points from each class. SVMs are a sort of selective and effective classifier that is frequently used for classification of huge data, perform sentiment analysis. It can handle extremely high dimensionality space challenges that are more difficult to overcome than regression barriers.

One of the most well-known supervised ML models is the SVM, which is used for prediction or classification. It determines feature space hyperplane which provides variation across labels or classes. An SVM model considers training data points as the main aspect in the feature space, which is mapped to other points from different classes that are separated in a feasible manner. Then, test data points are grouped and classified into two sets, based on which hyperplane those data points are plotted.

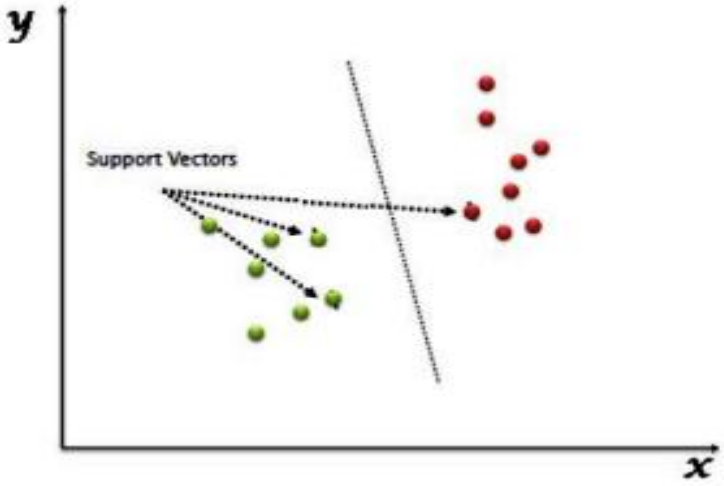

Figure 3:SVM hyperplane with data points.

It generates an ideal hyperplane that classifies new instances using labelled training data. By combining an SVM algorithm with a boosting technique, the authors achieved 88.1 percent accuracy, which is higher than using alone SVM. SVM learning, one of the proposed classification algorithms for heart disease classification, attained an accuracy of 88.1 percent, which is lower than ML accuracy.

Random Forest (RF)

One of the supervised classificationalgorithms[17] constructs forest with a specific number of trees. Accuracy of prediction using this algorithm is linearly relatedto the number of trees in the forest and by filling the missing data. The random forest classifierdoesnot overfit the model if there are more trees in the forest. Random forests are an ensemble learning technique that works by training by building a huge number of decision trees and then categorizing the outcomes based on the individual trees. In a decision tree, there is a danger of overfitting. The random forest's use of multiple trees [18] reduces the possibility of overfitting.

K-Nearest Neighbour (K-NN) Algorithm

The letter K stands for "nearest neighbour." It is used to get instance values in classification and regression problems by using a user-defined value as "k." Because it employs all the examples, it is a nonparametric classification approach [19] based on instances. It classifies an object based on the majority vote of its near neighbours. KNN is a simple and inherent classification techniques that is nevertheless quite complex.

KNN seeks to find points in training data set that are closest to the test data. Because it made no assumptions about the data, KNN [20]was frequently employed for classification without knowledge of the data distribution. To put it another way, the class of a training node will be predicted using a range of distance metrics, one of which may be a simple Euclidean distance.

$$D = \sqrt{\sum_{I=1}^{N} (X_I - Y_I)}$$

The researchers utilized a hybrid of KNN and achieved 96.7 percent accuracy in detecting cardiac issues. Four classification techniques were investigated, including RF. When the accuracy of DT, SVM, and KNN were compared, KNN was shown to be the most accurate. In this study, KNN algorithm is able to predict heart illness with 96.7 percent accuracy.

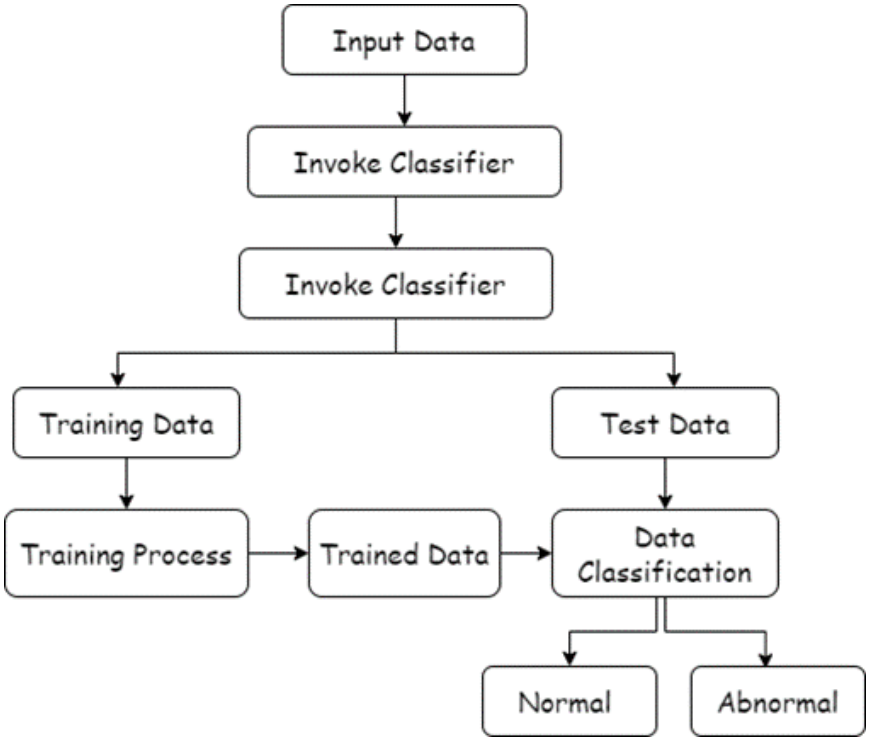

Figure 4: Overall Process for Heart Disease Prediction

After that, the classifier is used to assess the performance of each clustered dataset. Based on their low rate of error, the top performing models are chosen from the following findings. Several studies, as shown in the table below, employed well-known machine learning based classifiers for heart disease prediction, such as DT, SVM, KNN, and RF learning algorithms. They calculated performance indicators such as specificity, sensitivity, and accuracy in each blueprint.

4. Results and Discussion

A) Comparative Analysis

The classification of the ML algorithm, which is evaluated with various strategies [21] using the present method, increased the estimation of the suggested strategy in this area. The dataset is used to do the prediction without handling missing values does not produce promising results. The quality of dataset is improved by identifying the promising attributes which are highly correlated with the class label and by handling missing values. Only promising attributes with complete dataset is used for creating models. The proposed approach compares the results of above-mentioned algorithms using varied results analysis technique such as specificity, sensitivity, and classification accuracy.

B) Sensitivity, Specificityand Accuracy

In this research work, performance ofproposed machine learning techniques implemented are compared to that of existing approaches, and the proposed KNN classifier is compared to that of current SVM. Following equations are used to assess specificity, sensitivity, and efficiency of the algorithms used for prediction used in the system fordiagnosis of heart disease.

$$\text{Specificity} = \left(\frac{\text{TN}}{\text{TN}+\text{FP}}\right) * 100$$

$$\text{Accuracy} = \left(\frac{\text{TP}+\text{TN}}{\text{TP}+\text{FN}+\text{TN}+\text{FP}}\right) * 100$$

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP}+\text{FN}} * 100$$

where, TP refers thecountof True Positiveestimates, FP to False Positiveestimates, TN to True Negativeestimatesand FN to False Negativeestimates of the implemented system.

Parameters	KNN (%)	DT (%)	SVM (%)	RF (%)
Sensitivity	92	89.9	82	78.6
Specificity	94.6	86.5	81.9	66.4

Accuracy	96.7	89.6	88.1	79.8

Table 1: Performance analysis of algorithms used for prediction

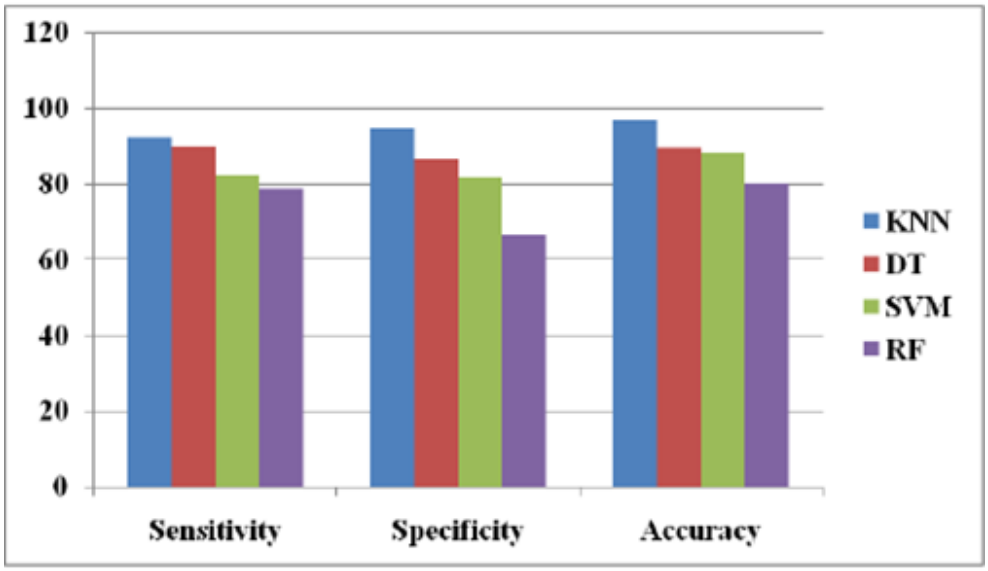

Figure 5: Classification of Heart Disease

Table 1 shows comparison results for the aforesaid ML approaches used in our system. As seen in the table above, the suggested KNN algorithm outperforms other machine learning models for the heart disease dataset used in the system. Even though decision tree, SVM, and RF classifiers producesadmissible results with their valid scores, KNN realises best scores with92 percent, 94.6 percent, and 96.7 percent for sensitivity, specificity, and accuracy respectively.Comparative results is shown in Figure 5 in graphical format for better understanding and visualization. Thus, the system finalises themodel created by KNN for prediction of cardiovascular disorder for any given input data of patients.

5. Conclusion

Predicting cardiac disease is anexigentglitch in thefield of medical diagnosis. The death rate, however, could be lowered if the infirmity is found in early stages. This article examines a variety of different study efforts and delves into the numerous categorization algorithms and approaches used by researchers to create accurate predictions. Several research publications revealed that certain classifiers, algorithms, or approaches, such as SVM, KNN, DT, and RF, were more accurate than others. New ensemble approaches that combine several classification algorithms, such as hybrid models or multiple learning models, produce better, more accurate results. According to the current study, among several classifiers implemented,KNNgive strong results with 96.7 percent accuracy.

Future predictions will be done only with model created with KNN. This system will be improved further by using principal component analysis [22]to reduce the features used for prediction.

References

[1]. S. Mohan, C. Thirumalai & G. Srivastava, "Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques", IEEE Access, vol.7, 2019. https://doi.org/10.1109/ACCESS.2019.2923707

[2]. K. Mathan, P. M. Kumar, G. Manogaran, P. Panchatcharam, and R. Varadharajan, " A novel Gini index decision tree data mining method with neural network classifiers for prediction of heart disease", Sprinjer, April 2018. https://doi.org/10.1007/s10617-018-9205-4

[3]. Min Chen, YixueHao, Kai Hwang, Fellow,Lu Wang, and Lin Wang, "Disease Prediction by Machine Learning over Big Data from Healthcare Communities", IEEE, vol. 15, pp. 215-227, 2017.https://doi.org/10.1109/ACCESS.2017.2694446

[4]. Azuaje, F. Witten IH, Frank E: Data Mining: Practical Machine Learning Tools and Techniques 2nd edition. BioMed Eng OnLine 5, 51 (2006). https://doi.org/10.1186%2F1475-925X-5-51

[5]. V. Ramalingam, A. Dandapath and M. Raja, "Heart Disease Prediction using Machine Learning Techniques: a survey", International Journal of Engineering and Technology [IJET], vol.7, pp.684–687, 2018.http://dx.doi.org/10.14419/ijet.v7i2.8.10557

[6]. KarenGárate-Escamila, E. Andrès, and A. E. Hassani, "Classification models for heart disease prediction using feature selection and PCA," Informatics in Medicine Unlocked, vol. 19, Article ID 100330, 2020. https://doi.org/10.1016/j.imu.2020.100330

[7]. Mr. ChalaBeyene, and Pooja Kamat, "Survey on Prediction and Analysis the Occurrence of Heart Disease Using Data Mining Technique", International Journal of Pure and Applied Mathematics, 2018.http://dx.doi.org/10.1109/ICEDSS.2018.8544333

[8]. C. B. C. Latha and S. C. Jeeva, "Improving the accuracy of prediction of heart disease risk based on ensemble classification techniques," Informatics in Medicine Unlocked, vol. 16, Article ID 100203, 2019. http://dx.doi.org/10.1016/j.imu.2019.100203

[9]. J. Singh, A. Kamra, H. Singh, "Prediction of Heart Diseases Using Associative Classification", IEEE, 2016.https://doi.org/10.1109/WECON.2016.7993480

[10]. K. Mathan, P. M. Kumar, P. Panchatcharam , G. Manogaran, R. Varadharajan, " A novel Gini index decision tree data mining method with neural network classifiers for prediction of heart disease", Sprinjer, April 2018.https://link.springer.com/article/10.1007/s10617-018-9205-4

[11]. Subbulakshmi, S., Nambiar, A.R., Arun, A.K., Al Faizi, F., Harish, V.N. (2022). Prediction of Priority to Individual for COVID Vaccine Distribution Using Soft Computing Techniques. In: Shakya, S., Du, KL., Haoxiang, W. (eds) Proceedings of Second International Conference on Sustainable Expert Systems . Lecture Notes in Networks and Systems, vol 351. Springer, Singapore. http://dx.doi.org/10.1007/978-981-16-7657-4_15

[12]. Fan C, Chen M, Wang X, Wang J and Huang B (2021) A Review on Data Preprocessing Techniques Toward Efficient and Reliable Knowledge Discovery from Building Operational Data. Front. Energy Res. 9:652801. doi: 10.3389/fenrg.2021.652801.http://dx.doi.org/10.3389/fenrg.2021.652801

[13]. Miao, Jianyu & Niu, Lingfeng. (2016). A Survey on Feature Selection. Procedia Computer Science. 91. 919-926. 10.1016/j.procs.2016.07.111.https://doi.org/10.1016/j.compeleceng.2013.11.024

[14]. Sarker, I.H. Machine Learning: Algorithms, Real-World Applications and Research Directions. SN COMPUT. SCI. 2, 160 (2021). https://link.springer.com/article/10.1007/s42979-021-00592-x

[15]. S. Subbulakshmi, A. Arjun and K. Jayaraj, "A QoS Value Prediction for Web Services Using Ethical User Identification," 2019 2nd International Conference on Intelligent Computing, Instrumentation and Control Technologies (ICICICT), 2019, pp. 1605-1609, doi: 10.1109/ICICICT46008.2019.8993366.https://doi.org/10.1109/ICICICT46008.2019.8993366

[16]. Subbulakshmi, S., Ramar, K., Omanakuttan, A., Sasidharan, A. (2019). Automated Analytical Model for Content Based Selection of Web Services. In: Thampi, S., Marques, O., Krishnan, S., Li, KC., Ciuonzo, D., Kolekar, M. (eds) Advances in Signal Processing and Intelligent Recognition Systems. SIRS 2018. Communications in Computer and Information Science, vol 968. Springer, Singapore. https://doi.org/10.1007/978-981-13-5758-9_26

[17]. A. Ajesh, Jayashree Nair, and Jijin, P. S., “A Random Forest Approach for Rating-based Recommender System”, in 2016 International Conference on Advances in Computing, Communications and Informatics (ICACCI), Jaipur, India, 2016 https://ieeexplore.ieee.org/document/7732225

[18]. R. Prasanna Kumar, “An empirical study on machine learning algorithms for heart disease prediction”, IAES International Journal of Artificial Intelligence, 2021.https://doi.org/10.11591/ijai.v11.i3.pp1066-1073

[19]. R. Kalyanasundaram, Prasanth, A., Tamizhselvan, B. R., and Kumaran U., “Calculating the heart disease in switzerland using Pearson's correlation”, in 2017 International Conference on Trends in Electronics and Informatics (ICEI), Tirunelveli, India, 2017.https://doi.org/10.1109/ICOEI.2017.8300885

[20]. K. Taunk, S. De, S. Verma and A. Swetapadma, "A Brief Review of Nearest Neighbor Algorithm for Learning and Classification," 2019 International Conference on Intelligent Computing and Control Systems (ICCS), 2019, pp. 1255-1260, doi: 10.1109/ICCS45141.2019.9065747.https://doi.org/10.1109/ICCS45141.2019.9065747

[21]. Ba'abbad, I., Althubiti, T., Alharbi, A., Alfarsi, K. and Rasheed, S. (2021) A Short Review of Classification Algorithms Accuracy for Data Prediction in Data Mining Applications. Journal of Data Analysis and Information Processing, 9, 162-174.https://doi.org/10.4236/jdaip.2021.93011

[22]. Guo, Qu & Wu, W &Massart, D.L &Boucon, C & de Jong, Sjoerd. (2002). Feature selection in principal component analysis of analytical data. Chemometrics and Intelligent Laboratory Systems. 61. 123-132. 10.1016/S0169-7439(01)00203-9.https://doi.org/10.1016/S0169-7439(01)00203-9